

## MODELAGEM DE EXPERIMENTOS PLANEJADOS COM RESPOSTAS DISCRETAS

Francisco Regilson SOUZA<sup>1</sup>  
Júlio Sílvio de Sousa BUENO FILHO<sup>2</sup>

- RESUMO: Neste trabalho se apresenta uma justificativa para utilização dos modelos lineares generalizados mistos (MLGM), como opção, em comparação ao modelo linear generalizado (MLG), para experimentos planejados com valores discretos para a variável resposta, quando se tomam  $m$  observações ( $m > 1$ ) em cada unidade experimental (UE). O MLGM se obtém adicionando um componente aleatório ao preditor linear, para capturar as variações existentes entre UE e comparando sua análise a do MLG em experimentos simulados com respostas discretas (da distribuição binomial ou Poisson). Foi considerado um arranjo experimental de UE em um delineamento inteiramente casualizado (DIC) e simulados experimentos supondo conhecidos os efeitos destas UE's. No MLGM as respostas dos tratamentos foram combinadas as das UE em um modelo linear. Foram simuladas respostas discretas e aplicadas às ligações canônicas para modelos binomiais e de Poisson. Os experimentos resultantes foram analisados das duas formas (MLG e MLGM). As análises foram feitas usando Software R 2.14 com 4.000 simulações para cada configuração, com diferentes valores de  $m$ . Em todos os parâmetros utilizados na comparação dos modelos os MLGM mostraram-se mais bem ajustados aos dados dos experimentos do que os MLG e devem ser utilizados em sua substituição nesses casos.
- PALAVRAS-CHAVE: Modelo binomial; modelo Poisson; modelos generalizados; modelos generalizados mistos.

### 1 Introdução

Os Modelos Lineares Generalizados (MLG - NELDER e WEDDERBURN, 1972) têm sido uma ferramenta muito utilizada na análise de dados em diferentes áreas. No entanto, para dados na forma de proporções e de contagens, frequentemente, as observações obtidas apresentam maior variabilidade do que é possível explicar pelos MLGs. Isto leva a potenciais diagnósticos de superdispersão em situações de experimentos planejados.

---

<sup>1</sup> Instituto Federal de Educação, Ciência e Tecnologia da Bahia - IFBA – Campus Jequié, Departamento de Ensino, CEP: 45.201-570, Jequié, BA, Brasil. E-mail: [regisouzza@hotmail.com](mailto:regisouzza@hotmail.com)

<sup>2</sup> Universidade Federal de Lavras - UFLA, Departamento de Ciências Exatas, Caixa Postal 3037, CEP: 37200-000, Lavras, MG, Brasil. E-mail: [jssbueno@dex.ufla.br](mailto:jssbueno@dex.ufla.br)

Uma das possíveis causas desse problema está na falta de um componente no preditor linear que descrevam efeitos aleatórios, o que leva a variável aleatória apresentar dispersão maior do que a predita pelos modelos Binomial ou Poisson, fenômeno este denominado de *superdispersão* (McCULLOCH e SEARLE, 2001).

Nos experimentos aleatorizados em que se medem variáveis aleatórias contínuas, é apenas necessário assumir que os efeitos de tratamento e de Unidades Experimentais sejam aditivos. Como resultado, segue-se que o modelo Gauss - Markov Normal (GMN) pode ser utilizado como boa aproximação da distribuição de referência sob hipótese de nulidade dos efeitos do modelo. Isto leva a algoritmos simples de modelos fixos para inferir sobre diferenças entre médias de tratamentos. Se em um experimento aleatorizado forem tomadas em cada unidade experimental, várias observações da variável de interesse, certamente a distribuição da variável resposta ainda seguirá o modelo GMN, porém os testes F para tratamentos serão realizados com a estimativa de variância entre unidades experimentais, ou seja, na realização do teste F para diferenças entre tratamentos a distribuição de referência (a ser usada no denominador do teste) é a das UEs e não a das Unidades Observacionais (UO).

Consideremos uma situação análoga em que se deseja analisar um experimento em um Delineamento Inteiramente Casualizado - DIC com  $r$  repetições por tratamento para uma variável com uma só resposta Bernoulli em cada unidade experimental. Neste caso, o modelo correto é o modelo binomial e o teste de razão de verossimilhanças (ou diferença na devianças) adequado comparará o modelo com uma só média geral ao modelo com uma média para cada tratamento. O procedimento é análogo a um modelo generalizado de efeitos fixos para tratamento – MLG (NELDER e WEDDERBURN, 1972).

A situação será diferente quando se tomam várias (digamos,  $m$ ) observações Bernoulli na unidade experimental. Cada UE tem uma resposta binomial com sua proporção característica. O teste adequado deveria levar em conta a variação das UE. O modelo usual (MLG) não o faz. Em lugar disso, a análise MLG equivale a descrever a variação em um DIC com  $r.m$  repetições para cada tratamento. Desta forma as diferentes UEs do mesmo tratamento associa-se apenas a variação devida à proporção única do tratamento. Tal análise pode levar a maiores taxas de erros do tipo I e às excessivas declarações de superdispersão (estimativas de deviance residual com valor muito superior ao esperado pelas distribuições estimadas para as médias de tratamento declaradas). Nestas situações, uma alternativa que vem ganhando destaque é a utilização dos modelos lineares generalizados mistos - MLGM (McCULLOCH e SEARLE, 2001).

O objetivo deste trabalho é comparar o ajuste de modelos generalizados (MLG) ao de modelos generalizados mistos (MLGM) na análise de experimentos planejados com respostas discretas. As comparações se darão em um estudo de simulação de experimentos com respostas que seguem a distribuição binomial e de Poisson.

## 2 Material e métodos

Nesta seção, inicialmente procura-se justificar a adoção dos MLGM recorrendo à derivação de um modelo linear, seguindo procedimento adotado por Hinkelmann & Kempthorne (2008) para variáveis contínuas, com a partição do erro experimental em componentes capazes de quantificar, de forma isolada, os efeitos entre e dentro de UE.

Para a realização deste trabalho, foi ajustada uma função do R para analisar um experimento planejado em um DIC com resposta binomial em cada UE, considerando que, a soma de  $m$  Unidades Observacionais - UO independentes do tipo Bernoulli (0,1) na UE, equivale a uma resposta binomial com tamanho amostral  $m$ .

Num experimento com resposta binomial em que cada tratamento  $t$  possui  $r$  repetições e cada uma das UE recebe  $m$  respostas Bernoulli independentes, a análise usual do MLG trata como  $r.m$  respostas Bernoulli independentes para cada tratamento, que corresponde a uma análise de dados com uma distribuição binomial com tamanho amostral  $r.m$ . Mas neste exemplo ocorreram apenas  $r$  repetições da resposta binomial para cada tratamento, ou seja, em cada uma das UE ocorre uma distribuição binomial com proporção distinta (para as suas  $m$  respostas Bernoulli independentes). Nesse caso existem diferentes distribuições entre e dentro de UE.

Quanto às variações existentes entre UO da mesma UE foi derivado um modelo linear associado a um experimento planejado em um DIC, que apresente um componente que seja capaz de quantificar de forma isolada as variações das UEs. Este método segue todo processo de derivação do modelo linear proposto por Hinkelmann & Kempthorne (2008) para variáveis contínuas. As variações aleatórias devidas a UE foram adicionadas ao preditor linear, produzindo um novo modelo, que doravante será designado de MLGM.

## 2.1 Modelo derivado linear para variáveis contínuas em um DIC

Na experimentação agrícola, em geral, os ensaios contém em suas parcelas mais de um indivíduo (ou mais de uma UO), denominado tamanho amostral, e o que se faz, normalmente, é analisar o total da parcela ou a média por parcela. Com esse procedimento, a variação (ou variância) entre indivíduos dentro de uma mesma parcela (UE) é negligenciada. Esta variação dá origem à chamada variância dentro das UEs e a variação residual é a variância entre as UEs.

Considere um experimento num DIC em que se dispõe de  $N = mrt$  UO, sendo  $m$  destas pertencentes a cada uma das  $n = rt$  UE. Seja  $k$  o contador das UE,  $k = 1, \dots, n = rt$ , em que  $r$  são as repetições do tratamento  $i$ , indicadas pelo contador  $j$ ,  $j = 1, \dots, r$ . Aleatoriza-se os rótulos dos tratamentos às  $k$  UE. Esse procedimento estabelece uma distribuição de referência para o erro entre UE caso não haja efeito de tratamento. O processo de aleatorização pode ser expresso matematicamente com a introdução de variáveis aleatórias de delineamento do tipo Bernoulli (0,1), como segue:

$$\delta_{ij}^k = \begin{cases} 1, & \text{se a UE } k \text{ recebeu a repetição } j \text{ do tratamento } i \\ 0, & \text{caso contrário.} \end{cases}$$

Pode-se supor que se o tratamento  $i$  é aplicado a UE  $k$ , a resposta conceitual em alguma escala contínua pode ser representada por  $T_{ik}$ , e pode ser expressa como função de componentes, a qual tem identidade representada por:

$$T_{ik} = \bar{T}_{..} + (\bar{T}_i - \bar{T}_{..}) + (\bar{T}_k - \bar{T}_{..}) + (T_{ik} - \bar{T}_i - \bar{T}_k + \bar{T}_{..}). \quad (1)$$

Nota-se que a soma de todos os componentes é igual à resposta  $T_{ik}$ . Vale salientar que, para variáveis contínuas, esta decomposição proporciona argumentos geométricos para usar a análise de variância como uma estatística resumo do experimento.

Nesta abordagem, os componentes  $(\bar{T}_{.k} - \bar{T}_{..})$  e  $(T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..})$  da expressão (1) representam uma partição da variação responsável pelo erro experimental, Wilk e Kempthorne (1955):

$$(T_{ik} - \bar{T}_{i.}) = (\bar{T}_{.k} - \bar{T}_{..}) + (T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..}). \quad (2)$$

Por força da ortogonalidade dos componentes, a soma das variações captadas individualmente é igual à variação observada em  $(T_{ik} - \bar{T}_{i.})$ , sendo que  $(\bar{T}_{.k} - \bar{T}_{..})$  quantifica o efeito do erro de unidade ou unitário, que representa o quanto uma UE difere da sua média e o componente:

$$(T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..}) = (T_{ik} - \bar{T}_{i.}) - (\bar{T}_{.k} - \bar{T}_{..}), \quad (3)$$

quantifica o efeito do erro técnico e representa o efeito da interação do  $i$ -ésimo tratamento com a  $k$ -ésima UE. No modelo linear clássico, por exemplo, admite-se que o efeito de tratamento e a UE são aditivos, assumindo-se que a variação captada por esse componente é aditiva ao longo de todas as UE's, nesse caso, o efeito da interação tratamento vezes UE será negligenciada.

Ao final do período experimental, uma observação é feita, a qual será denotada por  $y_{ij}$  que representa a repetição  $j$  do  $i$ -ésimo tratamento. Usando as variáveis aleatórias de delineamento  $\delta_{ij}^k$ , pode-se estabelecer a relação entre o valor observado  $y_{ij}$  e a resposta conceitual  $T_{ik}$ , como segue:

$$y_{ij} = \sum_k \delta_{ij}^k T_{ik}. \quad (4)$$

Isto significa que se a UE  $k$  recebeu a repetição  $j$  do tratamento  $i$ , então será observado  $T_{ik}$ . Se fizer na resposta conceitual  $T_{ik}$  expressa em (1):

$$\mu = \bar{T}_{..}, \quad t_i = (\bar{T}_{i.} - \bar{T}_{..}), \quad p_k = (\bar{T}_{.k} - \bar{T}_{..}) \quad e \quad v_{ik} = (T_{ik} - \bar{T}_{i.} - \bar{T}_{.k} + \bar{T}_{..}),$$

obtém-se o modelo  $T_{ik} = \mu + t_i + p_k + v_{ik}$ , que substituindo em (4):

$$y_{ij} = \sum_k \delta_{ij}^k (\mu + t_i + p_k + v_{ik}),$$

que finalmente se tem:

$$y_{ij} = \mu + t_i + \sum_k \delta_{ij}^k p_k + \sum_k \delta_{ij}^k v_{ik}, \quad (5)$$

constitui o modelo derivado linear em um DIC.

Opcionalmente pode-se fazer  $u_j = \sum_k \delta_{ij}^k p_k$ ; se a UE  $k$  recebeu a  $j$ -ésima repetição do  $i$ -ésimo tratamento, observa-se o desvio  $p_k$  e passa a ser a  $j$ -ésima contribuição relativa ao tratamento  $i$  para o componente  $u_j$  e de maneira análoga,  $\varepsilon_{ij} = \sum_k \delta_{ij}^k v_{ik}$ , o modelo deduzido será:

$$y_{ij} = \mu + t_i + u_j + \varepsilon_{ij}. \quad (6)$$

Com o processo de aleatorização imposto e aditividade desses componentes com o auxílio das variáveis aleatórias de delineamento  $\delta_{ij}^k$ , tem-se que  $u_j \sim \text{IID}(0, \sigma_u^2)$  e

$$\varepsilon_{ij} \sim \text{IID}(0, \sigma_\varepsilon^2) \text{ com } e_{ij} = u_j + \varepsilon_{ij}.$$

Hinkelmann e Kempthorne (2008) apresentam um procedimento completo e as propriedades decorrentes desse processo para variáveis contínuas.

No processo exposto acima, cada repetição  $j$  do tratamento  $i$  recebe o resultado da soma das  $m$  mensurações independentes de sucessos ou fracassos, que podem ser codificadas como 1 ou 0, respectivamente, que nesse caso possui distribuição Bernoulli. Dentro da UE, a soma das  $m$  Bernoullis independentes representa uma distribuição Binomial de tamanho amostral  $m$ . No caso das  $m$  observações serem providas de contagens, as  $j$  repetições do tratamento  $i$  representam contagens que são números discretos, cada uma delas seguindo a distribuição Poisson, com médias associadas à UE (e não apenas ao tratamento).

## 2.2 Modelo linear generalizado e modelo linear generalizado misto

No MLG o único componente aleatório são as variáveis respostas  $Y_1, Y_2, \dots, Y_n$ , independentes e provenientes da mesma distribuição de probabilidade. As variáveis explicativas, responsáveis pelas informações a serem analisadas e estimação dos parâmetros, dão origem a um vetor de preditores lineares:

$$\eta_i = \sum_{j=1}^p x_{ij} \beta_j = x_i^t \boldsymbol{\beta} \text{ ou } \boldsymbol{\eta} = \mathbf{X} \boldsymbol{\beta}. \quad (7)$$

em que, neste trabalho,  $\mathbf{X}$  representa a matriz do Delineamento Inteiramente Casualizado (DIC) com 5 tratamentos e 5 repetições por tratamento,  $\boldsymbol{\beta}$  é o vetor de parâmetros  $p \times 1$  e  $\boldsymbol{\eta}$  é o preditor linear das UEs, de dimensão  $n$ . Para essa abordagem, a análise é feita seguindo o modelo:

$$Y_{ij} \sim \text{B}(\pi_i, m) \text{ para distribuição binomial e } Y_{ij} \sim \text{P}(\mu_i, m) \text{ para Poisson.}$$

Ou seja, as observações seguem a distribuição com média associada apenas ao tratamento.

No modelo linear generalizado misto, acrescenta-se uma componente aleatória  $u_j$  que contabiliza as variações existentes entre as UE. Dessa forma o preditor linear passa a ser descrito por:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \text{ com } \mathbf{u} = [u_1, \dots, u_j]^t, \text{ e } u_j = \sum_k \delta_{ij}^k p_k, \text{ para } j, k = 1, \dots, n = rt, \quad (8)$$

sendo  $\mathbf{Z}$  uma matriz identidade de ordem  $n$  e  $\mathbf{u}$  um vetor de efeitos aleatórios de dimensão  $n$ , mesma de  $\mathbf{X}\boldsymbol{\beta}$  definida em (8). O modelo obtido por esse processo será:

$$Y_{ij} \sim B(\pi_{ij}, m) \text{ para distribuição binomial e } Y_{ij} \sim P(\mu_{ij}, m) \text{ para Poisson.}$$

Ou seja, as observações seguem a distribuição com média associada tanto ao tratamento quanto à UE.

Por analogia ao que foi apresentado para variáveis contínuas, quando se teria a aproximação normal para a análise paramétrica, no caso das distribuições binomial e Poisson, pode-se pensar que a variável contínua está ligada à resposta pela função de ligação. As respostas discretas podem ser desta forma associadas à variável conceitual contínua, produzindo um modelo derivado que não corresponde ao modelo generalizado usual (em que as proporções nas UEs dependeriam apenas do efeito de tratamento). A aproximação contínua que parece adequada é o modelo generalizado misto, considerando efeitos aleatórios de UE.

Para o modelo binomial a função de ligação entre o preditor linear e a probabilidade de sucesso da distribuição da variável resposta a ser simulada na parcela é a logística, para a qual, a proporção para cada combinação de tratamento e UE pode ser encontrada pela função inversa:

$$\boldsymbol{\pi} = \frac{\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})}{1 + \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u})}. \quad (9)$$

Tais valores de  $\boldsymbol{\pi}$  representam probabilidades de sucesso que foram utilizadas para simulação de respostas binomiais em todas as UEs com diferentes tamanhos amostrais ( $m$ ).

Desta forma, em cada UE uma proporção diferente surge associada às  $m$  respostas Bernoullis independentes. Foram simuladas configurações com  $m$  variando entre os seguintes valores 1, 5, 10, 15 e 20. Foram efetuadas 4.000 simulações para cada configuração.

Considerando que as respostas  $Y_i$  são independentes e modeladas por uma distribuição Poisson com distribuição de probabilidade  $P(\mu)$  de parâmetro  $\mu > 0$ , sua função de ligação canônica é a logarítmica dada por:

$$\boldsymbol{\eta} = \boldsymbol{\theta} = \ln(\boldsymbol{\mu}), \text{ com inversa dada por: } \boldsymbol{\mu} = \exp(\boldsymbol{\eta}). \quad (10)$$

A função de ligação entre o preditor linear e a média da distribuição da variável resposta a ser simulada na parcela, para cada combinação de tratamento e UE, é a logarítmica, para a qual, a média pode ser encontrada pela função inversa:

$$\boldsymbol{\mu} = \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}). \quad (11)$$

Desta forma, para cada UE uma média diferente surge associada às  $m$  respostas Poisson independentes. Tais valores de  $\boldsymbol{\mu}$  representam as médias que foram utilizadas

para a simulação das respostas Poisson em todas as UE, com diferentes tamanhos amostrais ( $m$ ). Foram simuladas configurações com  $m$  variando entre os seguintes valores:  $m = \{1, 2, 3\}$ .

### 2.3 Medidas da qualidade do ajuste

Foram usadas como medidas de ajustes dos dois modelos a *deviance*, o critério de informação de Akaike (1974) - *AIC*, o critério bayesiano de Schwarz (1978) - *BIC*, as taxas de erros tipo I, as taxas de declaração de superdispersão, as distribuições das estimativas de efeitos de tratamentos nos dois modelos e o erro quadrático médio das estimativas de efeitos de tratamentos.

A *deviance* é baseada na estatística de razão de verossimilhanças de Wilks e calculada como a diferença entre a log-verossimilhança do modelo completo e do modelo sob investigação  $D(\boldsymbol{\theta}; \mathbf{y}) = -2[l(\boldsymbol{\theta}; \mathbf{y}) - l(\hat{\boldsymbol{\theta}}; \mathbf{y})]$ .

Como a distribuição das *deviances* é aproximadamente qui-quadrado, deve-se compará-la com o número de graus de liberdade do resíduo. A análise MLG estima o erro experimental  $\hat{\sigma}_e^2$  com  $(n.m - p)$  graus de liberdade; sendo  $n$  = número de UEs,  $m$  = número de UO nas UEs e  $p$  = número de parâmetros e compara o valor  $(n.m - p)\hat{\sigma}_e^2 / \sigma_e^2$  com o quantil da distribuição qui-quadrado com  $(n.m - p)$  graus de liberdade. Para a análise MLGM, o erro experimental foi particionado em componentes  $\sigma_e^2 = \sigma_u^2 + \sigma_\varepsilon^2$  para capturar as variações (resíduos) entre ( $\sigma_u^2$ ) e dentro ( $\sigma_\varepsilon^2$ ) das UEs, sendo estes estimados por  $\hat{\sigma}_u^2$  com  $(n - p)$  graus de liberdade e  $\hat{\sigma}_\varepsilon^2$  com  $n.(m - 1)$  graus de liberdade, respectivamente, que por força da ortogonalidade dos componentes do modelo deduzido, tem-se  $(n - p) + n.(m - 1) = n.m - p$ .

O MLGM compara o valor  $n.(m - 1)\hat{\sigma}_\varepsilon^2 / \sigma_\varepsilon^2$  com o quantil da distribuição qui-quadrado com  $n.(m - 1)$  graus de liberdade, para verificação do ajuste. Desta forma o MLGM tem menor número de graus de liberdade para o erro experimental, número compatível com o que se observaria em uma variável de resposta contínua no mesmo experimento. A análise MLG superestima o número de observações independentes que constituíram o experimento. Ao se comparar deviances decorrentes destas duas formas de análise é preciso ressaltar que este número de graus de liberdade efetivo do MLGM é uma referência básica para o bom ajuste.

Alguns critérios de verificação do ajuste dos modelos serão utilizados nestas comparações. O critério de informação de Akaike (*AIC*) é dado por  $AIC = deviance + 2p$  e o critério bayesiano de Schwarz (*BIC*),  $BIC = deviance + p.\log n$ , em que  $p$  é o número de parâmetros e  $n$ , o número de observações.

A taxa de erro tipo I representa a probabilidade de um modelo acusar diferenças entre tratamentos quando de fato não existem, e será estimada simulando configurações com efeitos nulos de tratamento e calculando a proporção de análises que rejeitam a hipótese nula.

A taxa de declaração de superdispersão será calculada como a porcentagem das vezes em que a variância do componente aleatório do MLG é maior que a prevista pela distribuição do modelo em análise (binomial ou Poisson).

Outro aspecto registrado é o acerto nas estimativas de efeitos de tratamentos, apresentados através das distribuições dos efeitos de tratamentos e do Erro Quadrático Médio – EQM, dado por:  $EQM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$ , sendo  $\hat{\theta}$  o estimador de  $\theta$ , que pode ser reescrito (MOOD et al., 1974) por  $EQM(\hat{\theta}) = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$ , em que  $[E(\hat{\theta}) - \theta]$  representa o viés das estimativas dos efeitos de tratamentos e mede o quanto a estimativa se aproxima do parâmetro.

Serão apresentadas as distribuições das estimativas resultantes dos 4.000 experimentos simulados, apenas para fins de ilustração. Adicionalmente foram calculados os erros quadráticos médios das estimativas pelos dois métodos.

## 2.4 Material simulado nas análises MLG e MLGM

Foi simulada uma área experimental composta de 25 UEs com efeitos conhecidos, seguindo um gradiente bem definido. O croqui de campo com os valores das médias das UEs na escala do preditor é dado pela Figura 1:

-4	-3	-2	-1	0
-3	-2	-1	0	1
-2	-1	0	1	2
-1	0	1	2	3
0	1	2	3	4

Figura 1 – Croqui com efeitos de UE.

Foram simulados ainda cinco tratamentos com efeitos conhecidos  $\mathbf{T} = \{-2, -1, 0, 1, 2\}$  a serem designados às UEs, com 5 repetições por tratamento, de forma aleatória e representa a parte fixa  $\mathbf{X}\boldsymbol{\beta} = \mathbf{T}$  do preditor linear. O vetor  $\mathbf{u}$  contém os efeitos das UEs:  $\mathbf{u} = (-4, -3, -2, -1, 0, -3, -2, -1, 0, 1, -2, -1, 0, 1, 2, -1, 0, 1, 2, 3, 0, 1, 2, 3, 4)^t$  e representa a parte aleatória do modelo  $\mathbf{Z}\mathbf{u} = \mathbf{u}$ . Estes dois grupos de efeitos foram combinados em um modelo linear, ou seja:  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ . A escolha desses valores (tratamentos e UE) para essa combinação de efeitos se justifica pela necessidade de tornar evidente os efeitos entre e dentro das UEs, para a comparação das análises MLG e MLGM.

## 2.5 Software

As análises estatísticas referentes ao ajuste dos modelos bem como os gráficos ilustrativos serão realizados com o auxílio do *software* livre R versão 2.14.2 (R, 2011) utilizando a função `glmer()` do pacote `lme4`.



## 2.6 Rotinas de análise

Para a análise dos resultados de cada experimento foram empregados o modelo generalizado fixo com efeitos de tratamentos e o modelo generalizado misto, com efeito de tratamentos e de UE. O MLG foi ajustado com a função básica do R  $glm(resp \sim trat, family = Binomial)$ , para a análise com resposta binomial e  $glm(y \sim trat, family = Poisson)$  para a análise Poisson. Sendo  $resp$  uma matriz em que na primeira coluna vão os sucessos e na segunda os fracassos para a variável observada, e  $y$  é a resposta para a variável observada Poisson.

Para a análise MLGM foi construída uma fonte de variação “parcela” correspondente às UEs. A análise foi feita usando o pacote *lme4* do R e sua função  $glmer(resp \sim trat + (1|parcela), family = Binomial)$ , para o modelo binomial e a função  $glmer(y \sim trat + (1|parcela), family = Poisson)$ , para a Poisson. Nesta especificação, será ajustada uma componente da variância para a distribuição do preditor linear entre parcelas.

## 3 Resultados

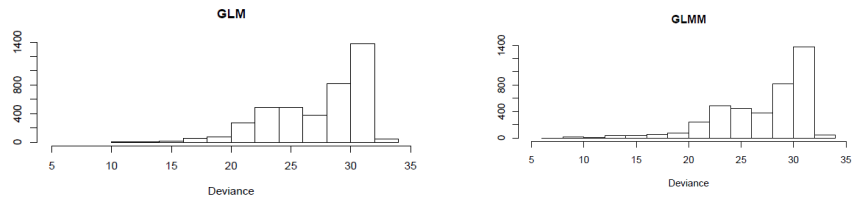
Os resultados para a distribuição binomial e Poisson serão apresentados de forma separada. Para cada distribuição serão apresentados a comparação dos resultados das análises MLG e MLGM, através das distribuições das estimativas das *deviances* residuais, taxas de erro tipo I, o critério de informação de Akaike - *AIC*, o critério bayesiano de Schwarz - *BIC*, taxas de declaração de superdispersão, distribuição das estimativas de efeitos de tratamentos e Erro Quadrático Médio - EQM das estimativas de efeitos de tratamentos. Estes são os parâmetros utilizados para o julgamento do ajuste e adequabilidade dos modelos.

### 3.1 Distribuição Binomial

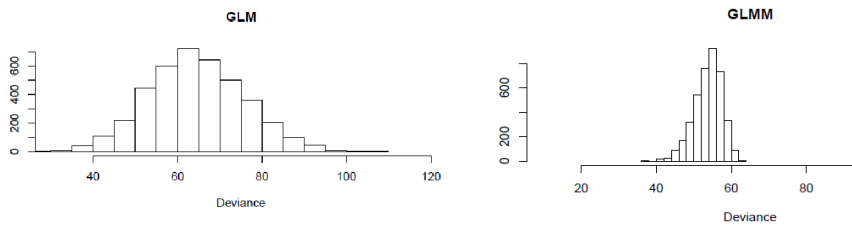
Nessa subseção, estão apresentados os resultados das estimativas das *deviances* residuais, taxas de erro tipo I, o critério de informação de Akaike - *AIC*, o critério bayesiano de Schwarz - *BIC*, taxas de declaração de superdispersão, distribuição das estimativas de efeitos de tratamentos e Erro Quadrático Médio - EQM das estimativas de efeitos de tratamentos nas duas análises MLG e MLGM referem-se aos dados com respostas binomiais.

#### 3.1.1 Distribuições das estimativas das *deviances* residuais

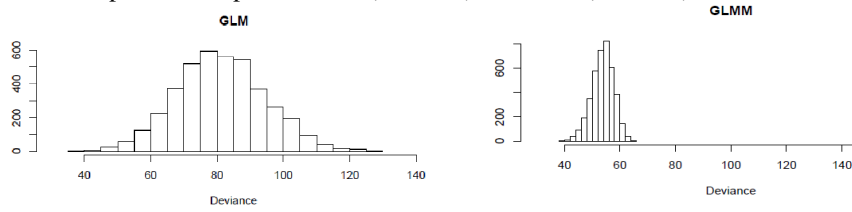
Na Figura 2, apresenta-se a comparação das distribuições das estimativas das *deviances* residuais dos dois modelos em análise MLG (gráfico à esquerda) e MLGM (gráfico à direita) com os respectivos graus de liberdade - GL do resíduo, para  $n = 25$  UEs,  $p = 5$  parâmetros e variando as UO dentro de cada UE  $m = 1, m = 5, m = 10, m = 15$  e  $m = 20$ .



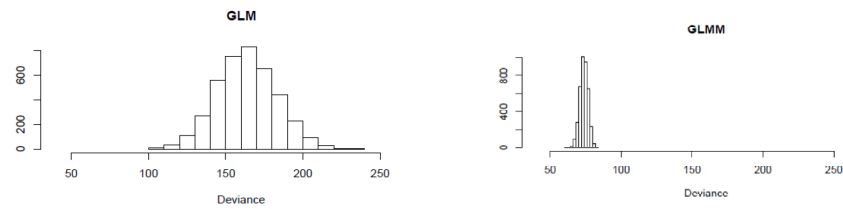
*Deviance para  $m = 1$  para o MLG (GL=20) e MLGM (GL=20).*



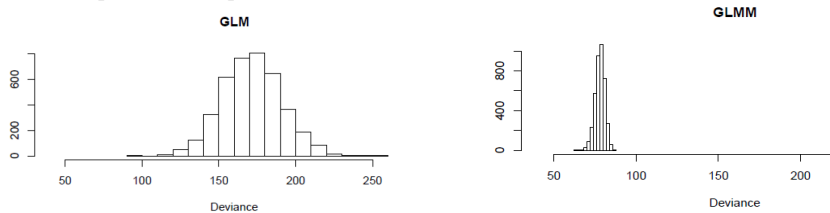
*Deviance para  $m = 5$  para o MLG (GL=120) e MLGM (GL=100).*



*Deviance para  $m = 10$  para o MLG (GL=245) e MLGM (GL=225).*



*Deviance para  $m = 15$  para o MLG (GL=370) e MLGM (GL=350).*



*Deviance para  $m=20$  para o MLG (GL=495) e MLGM (GL=475).*

Figura 2 - Distribuições das estimativas das *deviances* residuais dos modelos linear generalizado (MLG) e modelo linear generalizado misto (MLGM) com os tamanhos amostrais  $m = 1$ ,  $m = 5$ ,  $m = 10$ ,  $m = 15$  e  $m = 20$ .

Nota-se que não há diferença nas distribuições das estimativas das *deviances* residuais entre o MLG e o MLGM quando  $m = 1$ , como já era esperado, o que confirma a expressão da variância do componente aleatório do MLG

$$Var(y_{i.}) = m_i \mu (1 - \mu) \left[ 1 + \frac{m_i - 1}{2} \rho \right].$$

que, nesse caso, reduz-se a variância do modelo  $Var(y_{i.}) = \mu(1 - \mu)$ , uma vez que as repetidas observações dentro das UEs justificam parte da variação das UEs e na ausência de repetições, a análise pelos MLG e MLGM são equivalentes.

Obviamente neste caso, a pressuposição de distribuições Bernoullis independentes entre as UEs é válida (ou seja, há uma só distribuição binomial por tratamento). Pode-se notar também (pela expressão acima) que na medida em que os valores de  $m$  (UO) crescem ( $m = 5$ ,  $m = 10$ ,  $m = 15$  e  $m = 20$ ), a variância do componente aleatório cresce proporcionalmente, tornando-se maior que a variância do modelo, ocasionando aumento na *deviance*. Para os valores de  $m = 5$ ,  $m = 10$ ,  $m = 15$  e  $m = 20$ , nota-se que os valores das estimativas das *deviances* residuais para o MLGM são sempre menores que o do MLG. Além de menor média, as estimativas são mais concentradas, o que indica a adequação do modelo.

### 3.1.2 Taxas de erro tipo I

A Tabela 1 mostra as taxas de erro tipo I para os níveis de significância  $\alpha = 1\%$  e  $\alpha = 5\%$  para os dois modelos em análise MLG e MLGM com diferentes tamanhos amostrais  $m=1$ ,  $m=5$ ,  $m=10$  e  $m=15$ . A taxa de erro tipo I é um importante instrumento que auxilia na verificação da adequação ou não de um modelo, uma vez que as taxas de erro tipo I representam, em termos percentuais, a probabilidade de a análise acusar diferenças entre tratamentos que de fato não existem. É fácil perceber que o aumento no tamanho amostral  $m$  implica no crescimento das taxas de erro tipo I para o MLG, o contrário acontece no MLGM.

Tabela 1 - Taxas, em porcentagens, de erro tipo I nas análises MLG e MLGM, para o modelo binomial com diferentes tamanhos amostrais  $m=1$ ,  $m=5$ ,  $m=10$  e  $m=15$

$\alpha$	m=1		m=5		m=10		m=15	
	5%	1%	5%	1%	5%	1%	5%	1%
MLG	4,975	0,375	59,750	29,475	95,275	82,675	98,725	91,825
MLGM	4,375	0,975	4,425	0,200	0,775	0,000	0,025	0,000

\* MLG. Modelo Linear Generalizado; MLGM. Modelo Linear Generalizado Misto.

Para  $m = 20$  confirmam-se os resultados apresentados para as taxas de erro tipo I.

### 3.1.3 Taxas de declaração de superdispersão do MLG

Na Tabela 2, são apresentadas as taxas de declaração de superdispersão para o modelo MLG com diferentes tamanhos amostrais como  $m = 1$ ,  $m = 5$ ,  $m = 10$  e  $m = 15$ . A

taxa de declaração de superdispersão mostra, em termos percentuais, o quanto a variância do componente aleatório do MLG foi declarada maior que a do modelo em análise (binomial). No caso, trata-se de um estudo de simulação e podemos atribuir às altas taxas de erro tipo I observadas para o MLG a falta de um componente no preditor linear que incorpore ao modelo as variações entre UE.

Tabela 2 - Taxas, em porcentagens, de declaração de superdispersão, no MLG, para o modelo binomial com diferentes tamanhos amostrais  $m=1$ ,  $m=5$ ,  $m=10$  e  $m=15$

$\alpha$	m=1		m=5		m=10		m=15	
	5%	1%	5%	1%	5%	1%	5%	1%
MLG*	39,3	29,0	100	99,98	100	100	100	100

\* MLG. Modelo Linear Generalizado.

### 3.1.4 Medidas de ajuste AIC e BIC

Na Tabela 3 são mostradas as medidas de ajustes AIC e BIC para os dois modelos em análise MLG e MLGM com diferentes tamanhos amostrais como  $m=1$ ,  $m=5$ ,  $m=10$  e  $m=15$ .

Exceto para  $m=1$ , que nesse caso as análises são equivalentes, o MLGM apresentou menores valores para o AIC e BIC que o MLG.

Tabela 3 - Medidas de ajustes AIC e BIC, nas análises MLG e MLGM, para o modelo binomial com diferentes tamanhos amostrais  $m=1$ ,  $m=5$ ,  $m=10$  e  $m=15$

	m=1		m=5		m=10		m=15	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
MLG*	38,264	43,140	82,284	93,597	134,061	148,147	193,589	209,297
MLGM	38,236	43,112	63,393	74,706	77,272	91,357	86,823	102,530

\* MLG. Modelo Linear Generalizado; MLGM. Modelo Linear Generalizado Misto; AIC. Akaike; BIC. Bayesiano.

Na comparação entre dois modelos de análise, considera-se mais ajustado o que apresentar menor AIC e BIC, segundo esses critérios, o MLGM demonstra melhor ajuste que o MLG, com menor AIC e BIC, confirmando resultados anteriores.

### 3.1.5 Distribuição das estimativas dos efeitos de tratamentos

Na Figura 3, são apresentadas as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação  $\{-2, -1, 0, 1, 2\}$ . Nessa ordem, nos experimentos com  $m=5$  o viés das estimativas dos efeitos de tratamentos, dado por  $[E(\hat{\theta}) - \theta]$ , sendo o valor de  $\hat{\theta}$  a estimativa do parâmetro  $\theta$ , é uma medida da aproximação da estimativa do parâmetro e quanto menor é o viés, mais próxima é a estimativa do parâmetro e quando se tem  $E(\hat{\theta}) = \theta$ , o viés é nulo.

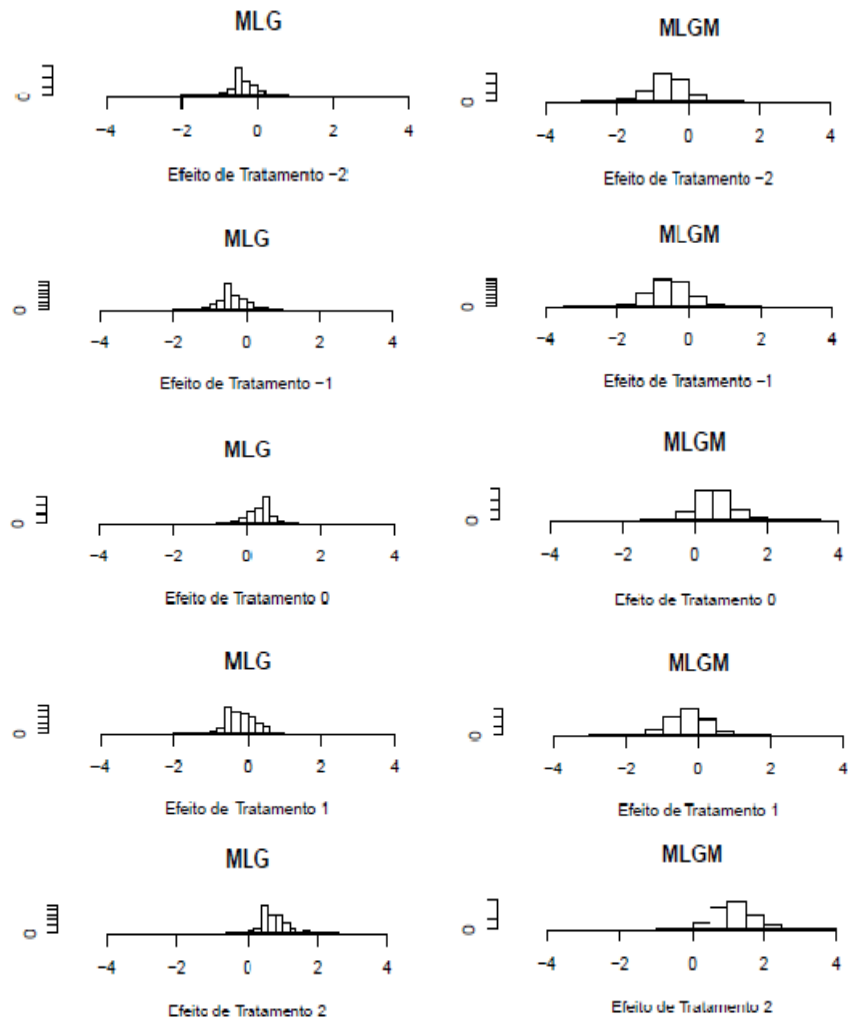


Figura 3 - Distribuições das estimativas de efeitos dos tratamentos  $\{-2, -1, 0, 1, 2\}$  nas análises do modelo linear generalizado misto (MLGM) e modelo linear generalizado (MLG), com  $m = 5$ .

Pode-se notar que o viés das estimativas do efeito de tratamento para  $m = 5$ , é menor no MLGM que no MLG, indicando que o MLGM se aproxima mais do verdadeiro efeito de tratamento e, conseqüentemente, apresenta melhor ajuste. Esta observação pode ser comprovada analiticamente e graficamente pelo Erro Quadrático Médio – EQM nos resultados que seguem.

Na Figura 4, são apresentadas as distribuições das estimativas de efeitos dos tratamentos que foram utilizados na simulação  $\{-2, -1, 0, 1, 2\}$  nos experimentos com tamanho amostral  $m=10$ . Nota-se que o viés  $[E(\hat{\theta})-\theta]$  das estimativas do efeito de tratamento é menor no MLGM do que no MLG, denotando que o MLGM é mais ajustado que o MLG, confirmando o resultado obtido para  $m=5$ .

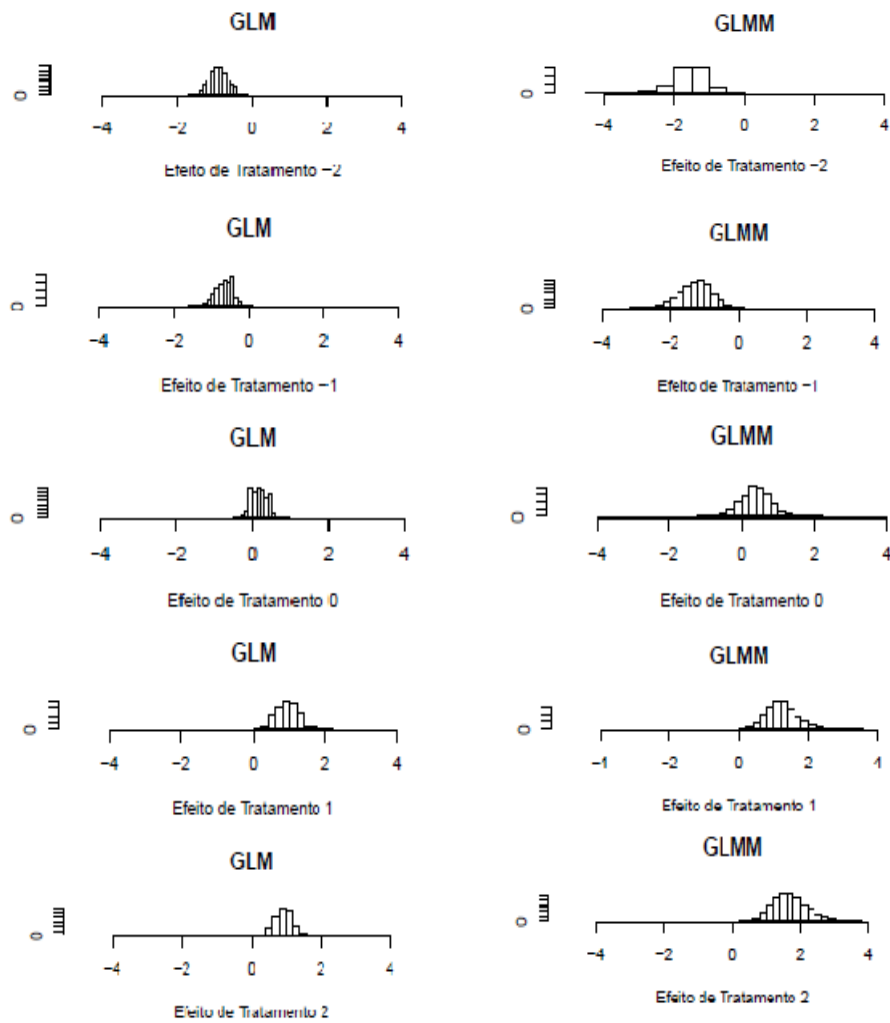


Figura 4 - Distribuições das estimativas de efeitos dos tratamentos  $\{-2, -1, 0, 1, 2\}$  nas análises de modelo linear generalizado misto (MLGM) e modelo linear generalizado (MLG), com  $m=10$ .

### 3.1.6. Distribuição do EQM das estimativas de efeitos de tratamentos

Na Figura 5, são apresentadas as distribuições do Erro Quadrático Médio – EQM das estimativas dos efeitos de tratamentos para os dois modelos de análise MLG e MLGM com os diferentes tamanhos amostrais  $m = 5$ ,  $m = 10$ ,  $m = 15$  e  $m = 20$ . O MLGM está apresentado no gráfico na cor vermelha e o MLG na cor preta. Pode-se notar que, em média, o EQM das estimativas de efeitos de tratamentos é menor para o MLGM que no MLG para todos os tamanhos amostrais considerados. Conforme esse critério, o MLGM aproxima mais do valor paramétrico que o MLG.

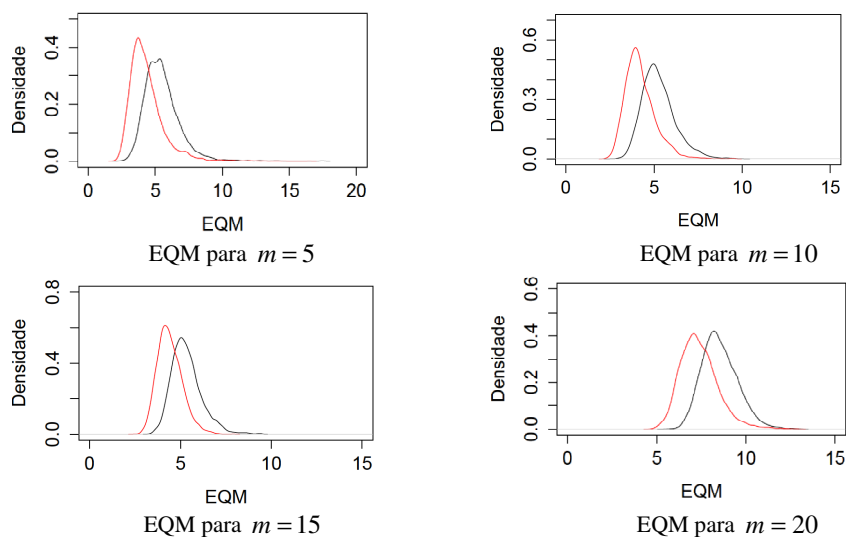


Figura 5 - Distribuições dos erros quadráticos médios - EQM no MLG (preto) e MLGM (vermelho) para  $m = 5, m = 10, m = 15, m = 20$  e  $m = 25$ .

### 3.2 Distribuição Poisson

Nessa subseção, estão apresentados os resultados das estimativas das *deviances* residuais, taxas de erro tipo I, o critério de informação de Akaike (1974) - *AIC*, o critério bayesiano de Schwarz (1978) - *BIC*, taxas de declaração de superdispersão, distribuição das estimativas de efeitos de tratamentos e Erro Quadrático Médio - EQM das estimativas de efeitos de tratamentos nas duas análises, MLG e MLGM referem-se aos dados com respostas Poisson.

### 3.2.1 Distribuições das estimativas das deviances residuais

Na Figura 6, são apresentadas as distribuições das estimativas das *deviances* residuais dos modelos MLG (gráfico à esquerda) e o MLGM (gráfico à direita), com os respectivos graus de liberdade do resíduo, para  $n = 25$  UEs,  $p = 5$  parâmetros e variando as UO dentro de cada UE,  $m = 1$ ,  $m = 2$  e  $m = 3$ . Nota-se que há diferença significativa entre as médias das estimativas das *deviances* residuais para os modelos em análise, sendo que o MLGM apresenta menor *deviance* residual que o MLG, para todos os tamanhos amostrais analisados.

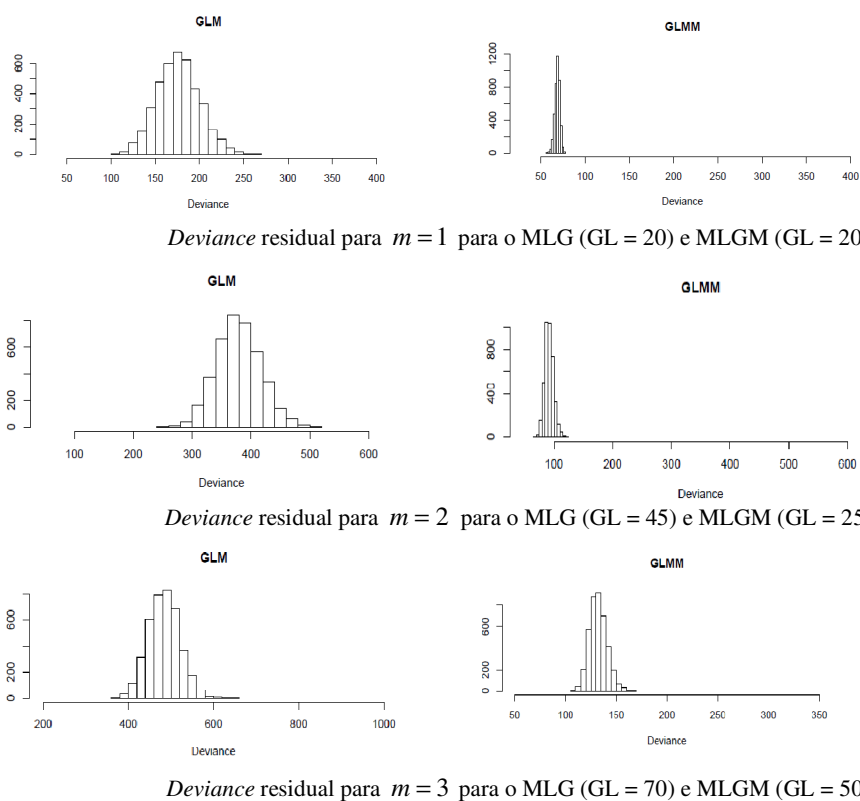


Figura 6 - Distribuições das estimativas das *deviances* residuais dos modelos linear generalizado (MLG) e linear generalizado misto (MLGM) com os tamanhos amostrais  $m = 1$ ,  $m = 2$  e  $m = 3$ .

É preciso ressaltar que a escala teve que ser reduzida para apresentar os histogramas do MLGM. Com base nisso, conclui-se que para todos os valores de  $m$ , inclusive  $m = 1$ , as deviances do MLGM são menores e mais concentradas, indicando sua melhor adequação. O aumento de  $m$  implica também em aumento na diferença entre as médias



das estimativas das deviances residuais dos dois modelos. Mesmo para  $m=1$ , o MLG é inadequado. Isto porque é mais correto modelar uma distribuição de Poisson com média diferente em cada UE. Espera-se que para valores de  $m$  muito maiores (talvez  $m > 15$ , por exemplo) um incremento rápido na aproximação pela análise normal (usando, por exemplo, uma transformação estabilizadora da variância). Isto deve ocorrer por um efeito de limite central ao se tomar médias de contagens (o mesmo efeito, embora presente na Binomial é mais lento por serem médias de variáveis com mesma distribuição Bernoulli).

### 3.2.2 Taxas de erro tipo I

Na Tabela 4, são apresentadas as taxas de erro tipo I para os dois modelos em análise MLG e MLGM para diferentes tamanhos amostrais como  $m=1$ ,  $m=2$  e  $m=3$  com níveis de significância  $\alpha=1\%$  e  $\alpha=5\%$ . Nota-se que o MLG apresenta taxa de erro tipo I de 100% para todos os tamanhos amostrais ( $m=1$ ,  $m=2$  e  $m=3$ ), o que indica, com probabilidade de 100%, acusar diferenças entre tratamentos que de fato não existe. O contrário acontece com o MLGM, com taxas bem menores, que reflete um modelo melhor ajustado, com resultados mais confiáveis para essa situação.

Tabela 4 - Taxas, em porcentagens, de erro tipo I para as análises MLG e MLGM, para o modelo Poisson com diferentes tamanhos amostrais  $m=1$ ,  $m=2$  e  $m=3$

$\alpha$	m=1		m=2		m=3	
	5%	1%	5%	1%	5%	1%
MLG*	100	100	100	100	100	100
MLGM	1,725	0,150	0,150	0,025	0,00	0,00

\* MLG. Modelo Linear Generalizado; MLGM. Modelo Linear Generalizado Misto.

### 3.2.3 Taxas de declaração de superdispersão do MLG

Na Tabela 5, são apresentadas as taxas de declaração de superdispersão para o modelo MLG com diferentes tamanhos amostrais  $m=1$ ,  $m=2$  e  $m=3$ , com níveis de significância  $\alpha=1\%$  e  $\alpha=5\%$ . Analogamente ao que acontece na análise binomial, o MLGM é modelado para não apresentar superdispersão. A dispersão declarada para a UE na análise do MLGM é o parâmetro canônico do modelo Poisson  $\sigma_u^2 = 1$ . As taxas de declaração de superdispersão representa, em termos percentuais, o quanto a variância do componente aleatório do modelo MLG é maior que a do modelo Poisson.

Tabela 5 - Taxas, em porcentagens, de declaração de superdispersão, nas análises MLG, para o modelo Poisson com diferentes tamanhos amostrais  $m=1$ ,  $m=2$  e  $m=3$

$\alpha$	m=1		m=2		m=3	
	5%	1%	5%	1%	5%	1%
MLG*	100	100	100	100	100	100

\* MLG. Modelo Linear Generalizado.

O MLG apresenta taxas de declaração de superdispersão de 100% para todos os tamanhos amostrais ( $m=1$ ,  $m=2$  e  $m=3$ ) e para todos os níveis de significância  $\alpha=1\%$  e  $\alpha=5\%$ , o que indica que a variância do componente aleatório do MLG é maior que a do modelo Poisson, necessitando de um componente no preditor linear para acomodar essa variabilidade extra e refletir a falta de ajuste do MLG para esse tipo de análise.

### 3.2.4 Medidas de ajuste AIC e BIC

Na Tabela 6, são apresentadas as medidas de ajustes AIC e BIC para os dois modelos em análise MLG e MLGM para diferentes tamanhos amostrais como  $m=1$ ,  $m=2$  e  $m=3$ . Na comparação de dois modelos de análise, será considerado mais ajustado o que apresentar menor valor para o AIC e BIC. Para todos os tamanhos amostrais ( $m=1$ ,  $m=2$  e  $m=3$ ) o MLGM apresenta menor AIC e BIC que o MLG, portanto, se ajusta melhor aos dados para esta situação.

Tabela 6 - Medidas de ajustes AIC e BIC, nas análises MLG e MLGM, para o modelo Poisson com diferentes tamanhos amostrais  $m=1$ ,  $m=2$  e  $m=3$

	m=1		m=2		m=3	
	AIC	BIC	AIC	BIC	AIC	BIC
MLG*	273,862	278,738	529,557	537,205	849,458	858,728
MLGM	75,214	80,090	111,930	119,577	142,522	151,792

\* MLG. Modelo Linear Generalizado; MLGM. Modelo Linear Generalizado Misto; AIC. Akaike; BIC. Bayesiano.

Na Tabela 6 pode-se notar, ainda, que na medida em que aumenta o tamanho amostral ( $m=1$ ,  $m=2$  e  $m=3$ ), piora o ajuste do MLG aumentando os valores do AIC e BIC, ampliando ainda mais a diferença para o MLGM.

### 3.2.5 Distribuição das estimativas dos efeitos de tratamentos

Na Figura 7 são apresentadas as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação  $\{-2, -1, 0, 1, 2\}$ , nessa ordem, nos experimentos, com tamanho amostral  $m=1$ . Nota-se que o viés  $[E(\hat{\theta})-\theta]$  das estimativas de efeito de tratamento é menor no MLGM do que no MLG, que sugere melhor ajuste do MLGM que o MLG nesse tipo de análise.

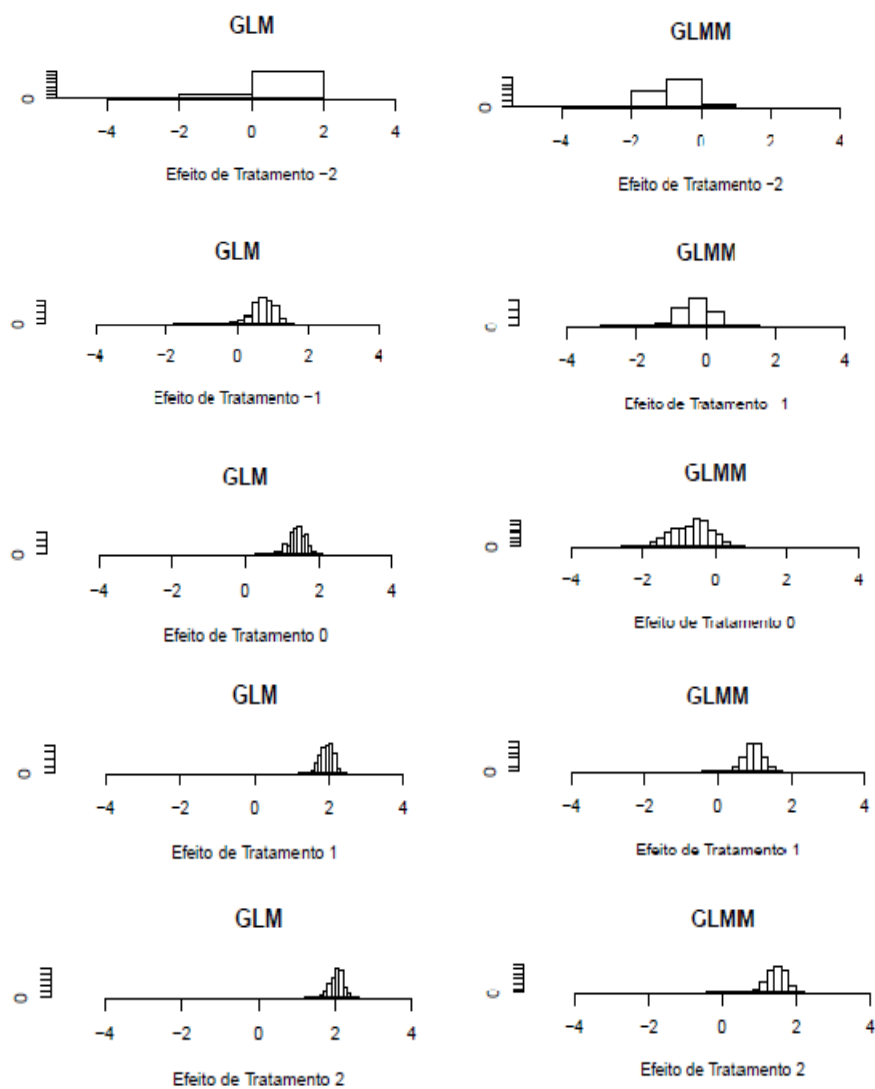


Figura 7 - Distribuições das estimativas de efeitos dos tratamentos  $\{-2, -1, 0, 1, 2\}$  nas análises MLGM e MLG com  $m = 1$ .

Na Figura 8, são apresentadas as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação  $\{-2, -1, 0, 1, 2\}$  nos experimentos com tamanho amostral  $m = 2$ . Confirmam-se os resultados anteriores para  $m = 1$ , em que o viés das estimativas de efeitos de tratamentos  $[E(\hat{\theta}) - \theta]$  é menor no MLGM do que no MLG, configurando melhor ajuste do MLGM que no MLG.

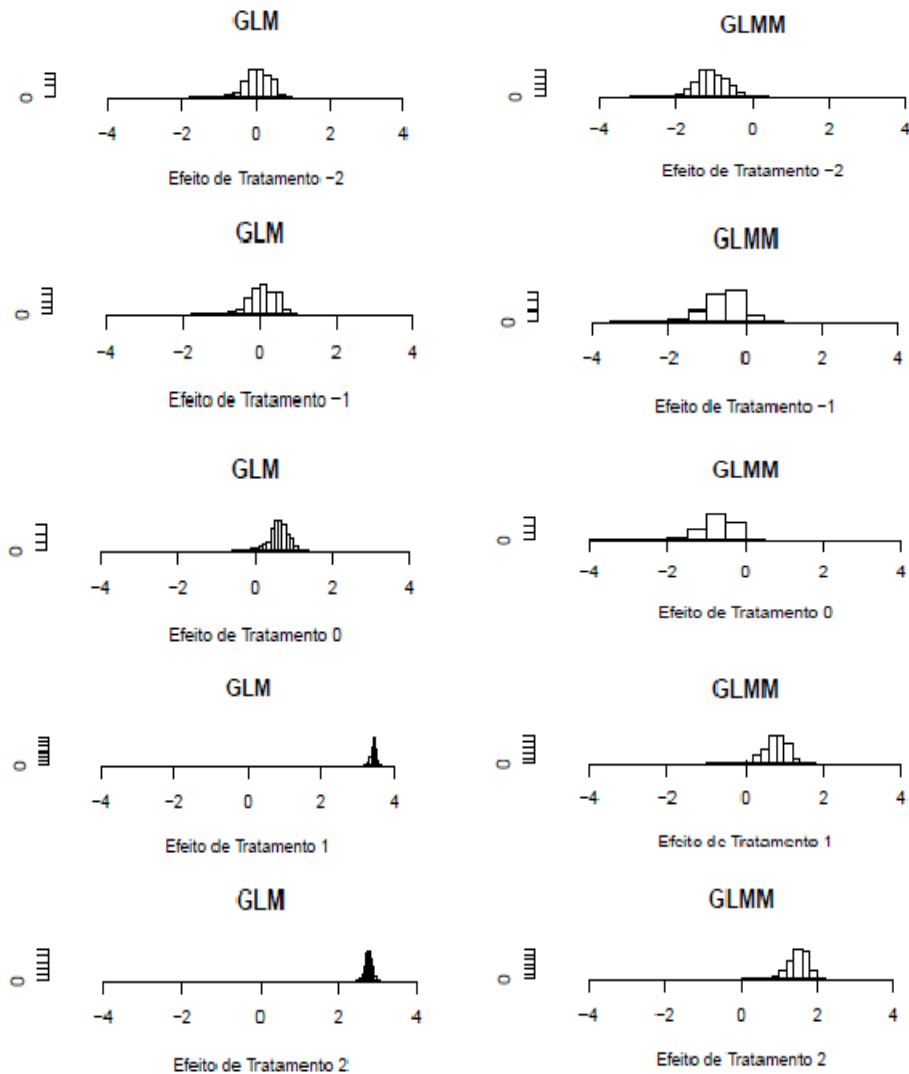


Figura 8 - Distribuições das estimativas de efeitos dos tratamentos  $\{-2, -1, 0, 1, 2\}$  nas análises MLGM e MLG, com  $m = 2$ .

Resultados também semelhantes aos anteriores ( $m = 1$  e  $m = 2$ ) são apresentados na Figura 9, para as distribuições das estimativas de efeitos dos tratamentos utilizados na simulação  $\{-2, -1, 0, 1, 2\}$  nos experimentos com tamanho amostral com  $m = 3$ . Nota-se que o viés das estimativas dos efeitos de tratamentos  $[E(\hat{\theta}) - \theta]$  é menor no MLGM do que no MLG, indicando melhor ajuste do MLGM que no MLG, para este tipo de análise.

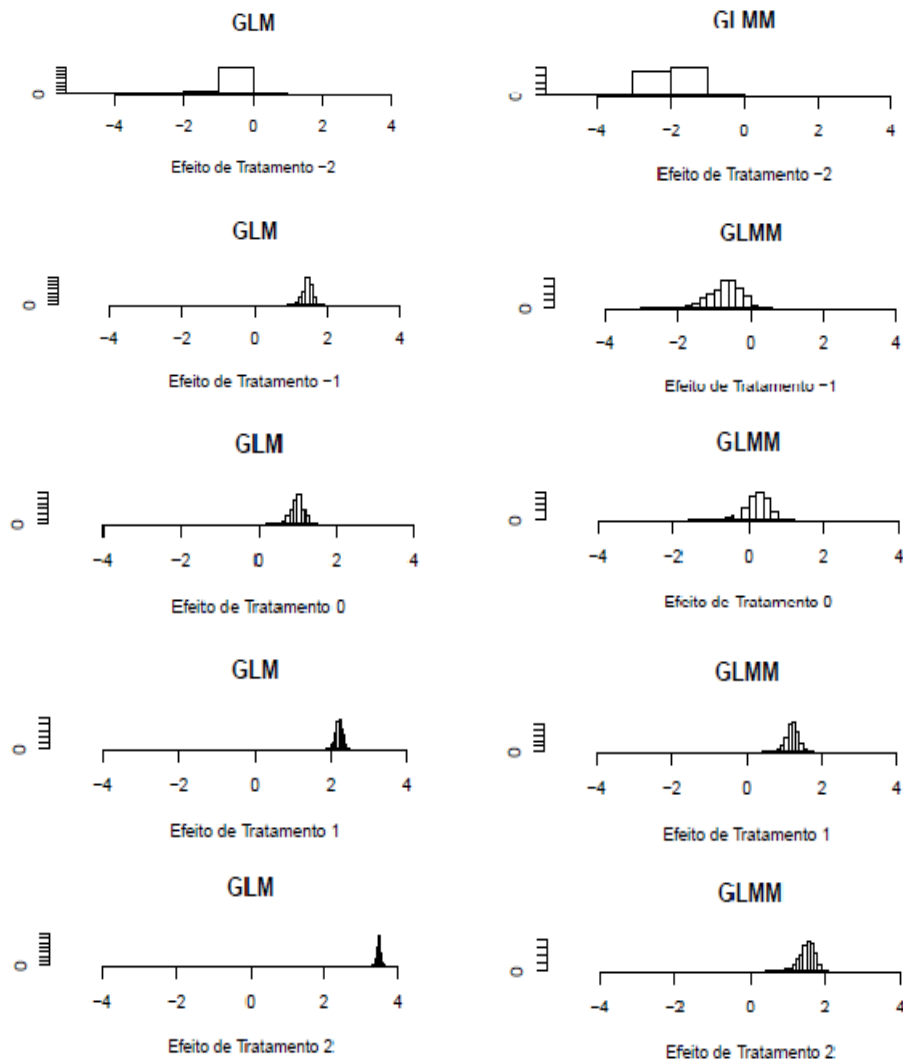


Figura 9 - Distribuições das estimativas de efeitos dos tratamentos  $\{-2, -1, 0, 1, 2\}$  nas análises MLGM e MLG, com  $m = 3$ .

### 3.2.4 Distribuição do EQM das estimativas de efeitos de tratamentos

Na Figura 10, são apresentadas as distribuições do Erro Quadrático Médio – EQM das estimativas dos efeitos de tratamentos para os dois modelos de análise MLG e MLGM com os diferentes tamanhos amostrais  $m = 1, m = 2$  e  $m = 3$ . O MLGM está apresentado no gráfico na cor vermelha e o MLG na cor preta. Pode-se notar que, em média, o EQM

das estimativas de efeitos de tratamentos é menor para o MLGM que no MLG, para todos os tamanhos amostrais considerados. Conforme esse critério, o MLGM aproxima mais do valor paramétrico que o MLG. Note que ao contrário do que se observou no modelo binomial, a diferença entre os EQM não aumenta “sempre” com o aumento do tamanho amostral  $m$ .

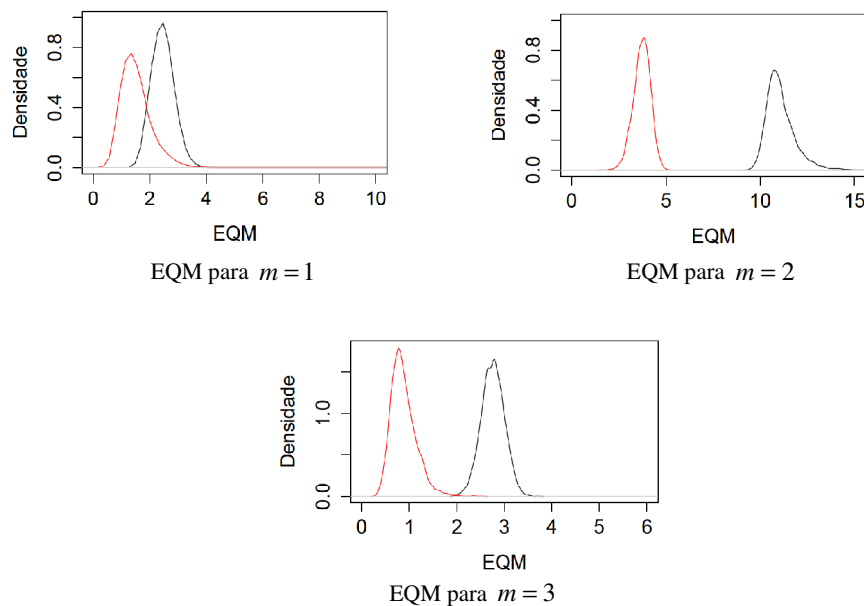


Figura 10 - Distribuições do EQM no MLG (preto) e MLGM (vermelho) para  $m=1$ ,  $m=2$ , e  $m=3$ .

#### 4 Discussão

As simulações foram capazes de ilustrar que quando as UEs recebem apenas uma resposta Bernoulli, as duas análises são equivalentes (Figura 2), como já era esperado, pois, nesse caso, não há variações da UE, apenas entre, que é a forma usual de análise MLG.

O modelo generalizado misto (MLGM) apresenta menor média das estimativas das deviances residuais, para a análise de ensaios binomiais do que o modelo generalizado fixo (MLG) para tamanhos amostrais das UEs maiores que 1 ( $m > 1$ ), ou seja, quando cada UE possui mais de uma resposta Bernoulli (Figura 2), mostrando-se, portanto, melhor ajuste. O mesmo ocorre com as demais medidas de ajustes *AIC* e *BIC* (Tabela 3), apresentando valores bem menores para o MLGM que o MLG, para todos os tamanhos amostrais maiores que 1 ( $m=5$ ,  $m=10$  e  $m=15$ ). No caso do *AIC*, as diferenças do MLG para o MLGM variam, em termos percentuais, de 29,8% (menor diferença) a 122,9% (maior diferença) a mais para o MLG. E no *BIC* essas diferenças do MLG para o MLGM, variam em termos percentuais, de 25,3% a 104,1% a mais para o MLG.

Na Tabela 1, mostra-se que para todos os tamanhos amostrais (inclusive  $m = 1$ ) as taxas de erro tipo I são muito maiores no MLG que no MLGM. Sendo a maior taxa do MLGM 4,4%, enquanto o MLG chega a uma taxa de 98,7% o que sugere, nesse caso, altíssima probabilidade 0,987 de acusar diferenças não existentes entre tratamentos, associada ao diagnóstico de superdispersão. Mais uma vez constata-se a maior adequabilidade do MLGM para dados com resposta binomial.

Na Tabela 2, confirmam-se os resultados anteriores quando apresenta a menor taxa de superdispersão em 39,3% para  $m = 1$  e 100% para  $m = 5$  no MLG. Isto indica um modelo superdisperso, mal ajustado e pode levar a conclusões equivocadas quanto às diferenças entre médias de tratamentos.

Para as estimativas dos efeitos de tratamentos (Figuras 3 e 4) o modelo MLG tende a apresentar estimativas mais agregadas de efeitos de tratamentos, no entanto erra com maior frequência o valor paramétrico dos efeitos, comprovado também através da distribuição dos EQMs. Isto confirma as maiores taxas de erro tipo I do MLG.

Como visto o MLGM aproximou-se bastante dos valores paramétricos dos efeitos simulados de tratamentos, gerando, no entanto, estimativas mais dispersas. Isso se justifica por esse modelo levar em conta variação entre UE. Ao contrário do MLG em que as UOs são todas consideradas independentes, o MLGM computa uma componente da variância que reflete a agregação nas UEs. Isto torna potenciais testes de média mais confiáveis e menos sujeitos a erros do tipo I.

O erro quadrático médio é uma medida resumo da precisão (acurácia) das estimativas de efeito de tratamento. Na Figura 5, tem-se a distribuição dos EQMs e observa-se que, em média, o MLGM se mostrou mais preciso que o MLG, para todos os tamanhos amostrais  $m > 1$  em relação ao MLG.

A Figura 6 apresenta as distribuições das estimativas das deviances residuais para os tamanhos amostrais  $m = 1$ ,  $m = 2$  e  $m = 3$ . Nota-se que a deviance média é muito menor no MLGM que no MLG. O mesmo ocorre com as demais medidas de ajustes *AIC* e *BIC* apresentando valores bem menores para o MLGM que o MLG para todos os tamanhos amostrais. No caso do *AIC*, as diferenças do MLG para o MLGM (Tabela 6) variam, em termos percentuais, de 264,12% até 496,02% a mais para o MLG. E no *BIC*, essas diferenças do MLG para o MLGM variam, em termos percentuais, de 248,06% até 465,72% a mais para o MLG. Denota-se, portanto, que o MLGM é mais ajustado para a análise de dados de contagens que a análise usual do MLG. Isto se justifica pela existência de variações entre e dentro das UEs.

Para todos os tamanhos amostrais, o MLG apresenta elevada declaração de falsos positivos, sendo a maior taxa apresentada pelo MLGM de 1,7%. Também aqui, a análise MLGM confirma os resultados anteriores, demonstrando mais ajustado para a análise de dados com resposta Poisson. Analogamente para a superdispersão o MLG foi muito propenso a declarar superdispersão ao contrário do MLGM.

As estimativas obtidas com o MLGM aproximaram-se bastante dos valores paramétricos dos efeitos simulados de tratamentos. Deve-se notar, no entanto, que as estimativas mais dispersas que as do MLG. Isto se justifica por que no MLG as UOs são todas consideradas independentes, já o MLGM computa uma componente da variância que reflete a agregação nas UEs dos mesmos tratamentos. Consideramos esta saída mais realista.

Considerando os resultados das duas análises MLG e MLGM e os parâmetros utilizados na comparação: *Deviance*, *AIC* e *BIC*, taxas de erro tipo I e de superdispersão e estimativas dos valores paramétricos e EQM, pode-se concluir que o MLGM, é mais ajustado e apresenta resultados mais confiáveis, para a análise de dados discretos que o MLG, para analisar efeitos fixos de tratamentos.

## Conclusões

Em experimentos planejados, a análise de modelos Bernoulli ou Binomiais utilizando a metodologia de modelos lineares generalizados mistos (MLGM) deve ser recomendada em substituição à análise de modelos generalizados fixos (MLG), sempre que se tiver mais que uma unidade observacional por unidade experimental.

Para os experimentos com resposta Poisson a conclusão é ainda mais forte: modelos generalizados mistos (MLGM) são mais adequados para a análise do que o modelo generalizado fixo (MLG) ainda que haja apenas uma unidade observacional por unidade experimental.

Em ambos os casos, ao contrário dos GLMs, os GLMMs permitem controlar a taxa de erro tipo I e não sofrem com diagnósticos de superdispersão.

## Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão de bolsas de doutorado e apoio a projeto de pesquisa.

À FAPEMIG pelo apoio aos autores.

SOUZA, F. R.; BUENO FILHO, J. S. S. Modelling planned experiments with discrete responses. *Rev. Bras. Biom.*, São Paulo, v.32, n.1, p.01-27, 2014

- *ABSTRACT: We present in this work a justification for the use of Mixed Generalized Linear Models (MGLM) as an option in comparison to the Generalized Linear Models (GLM) for experiments planned with discrete values for the response variables, when taking  $m$  observations ( $m > 1$ ) in each experimental unit (EU). The MGLM is obtained adding a random component to the linear predictor, in order to capture the existing variations between EU and comparing its analysis to that of the GLM in experiments simulated with discrete responses (of binomial or Poisson distributions). We considered the completely randomized design experimental arrangement and simulate experiments supposing the EU effects as known. In the MGLM, the responses of the treatments were combined with those from the EU in a linear model. We simulated discrete responses and used the canonic links from the binomial and Poisson models. The resulting experiments were analyzed in two manners (GLM and MGLM). The analyses were performed using the R 2.14 Software with 4000 simulations for each configuration, with different values for  $m$ . In all parameters used in the comparison between models, the MGLM showed most well-adjusted to the experimental data than those from the GLM and should be used in substitution in these cases.*
- *KEYWORDS: Binomial model; Poisson model; generalized models; generalized mixed models.*



## Referências

- AKAIKE, H. A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, Boston, v.19, n.6, p.716-723, 1974.
- HINKELMANN, K.; KEMPTHORNE, O. *Design and Analysis of Experiments*. 2.ed. New Jersey: John Wiley & Sons. Inc., 2008. 631p.
- McCULLOCH, C. E.; SEARLE, S. R. *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons. Inc., 2001. 325p.
- MOOD, A. L. M.; GRAYBILL, F. A.; BOES, D. C. *Introduction To The Theory Of Statistics*. 3.ed. Tokyo: McGraw-Hill, 1974. 564p.
- NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, London, v.135, n.3, p.370-384, 1972.
- R Development Core Team. R: *A language and environment for statistical computing*. R foundation for Statistical Computing, Vienna, Austria. 2011.
- SCHWARZ, G. Estimating the dimension of a model. *The Annals of Statistics*, Hayward, v.6, n.2, p. 461-464, 1978.
- WILK, M. B.; KEMPTHORNE, O. Fixed, Mixed and Random Models. *Journal of the American Statistical Association*, New York, v.50, p.1144-1167, 1955.

Recebido em 04.12.2013

Aprovado após revisão em 03.04.2014

## APÊNDICE

### Rotinas de Simulação

#### Binomial

A função “`analisa.Binomial(n,m,p,t)`” abaixo ao mesmo tempo gera e analisa o experimento. Os parâmetros desta função são: `n` (número de observações do experimento), `m` (número de observações da UE), `p` (média paramétrica, soma de média da unidade experimental e efeito de tratamento, na escala da proporção) e `t` (vetor com os rótulos de tratamentos sorteados). Utilizamos em “`analisa...`” a função `glmer()` do pacote `lme4` do R. A função apresentada neste exemplo devolve os seguintes resumos (respectivamente e para os modelos MLG e MLGM) *deviance* residual do modelo (`dmlg`, `dmlgm`) e coeficientes do preditor linear (`fmlg`, `fmlgm`). Outros resumos podem ser solicitados de forma análoga.

```
library(lme4)

analisa.Binomial <- function(n,m,p,t){
  y <- rbinom(n,m,p) # gera sucessos
  resp <- cbind(y, m-y) # concatena sucessos e fracassos
  Trat <- factor(t) # transforma t em fator
  mlg <- glm(resp ~ -1+Trat, family=binomial) # MLG
  dmlg <- deviance(mlg) # Deviance
  fmlg <- coef(mlg) # Coeficientes
  parc <- factor(1:n) # cria fator para as UE

# Análise MLGM

  mlgm <- glmer(resp ~ -1+Trat +(1|parc),family=binomial)
  dmlgm <- deviance(mlgm) # Deviance
  fmlgm <- fixef(mlgm) # Coeficientes
  return(list(dmlg, dmlgm, fmlg, fmlgm)) # saida
}
```

A forma de criar e passar as condições experimentais de um experimento para a função é:

```
baseTrat <- c(-2:2) # base para os efeitos
r <- 5 # número de repetições do DIC
t <- sample(rep(baseTrat,r)) # aleatoriza os tratamentos
n <- length(t) # tamanho do experimento
m <- 5 # ensaios Bernoulli por parcela
U <- c(c(-4:0),c(-3:1),c(-2:2),c(-1:3),c(0:4)) # efeitos
# de UE

eta <- t+U # preditor linear
p <- exp(eta)/(1+exp(eta)) # probabilidade de sucesso
```

## Poisson

A função “analisa.Poisson(n,m,p,t)” é análoga à da distribuição binomial, mas a variável resposta não precisa estar em duas colunas... O que muda nos parâmetros é que chamamos lambda o preditor. Utilizamos em “analisa” a função glmer() do pacote lme4 do R. Em negrito estão evidenciadas as alterações em relação à função anterior para a binomial.

```
analisa <- function(n,m,lambda,t){
  y      <- rpois(n,lambda) # amostra vetor Poisson(lambda)
  Trat   <- factor(t)
  mlg   <- glm(y ~ -1+Trat, family=poisson)
  dmlg   <- deviance(mlg)
  fmlg   <- coef(mlg)
  parc   <- factor(kronecker(1:length(t),rep(1,m)))
  mlgm  <- glmer(y ~ -1+Trat + (1|parc), family=poisson)
  dmlgm  <- deviance(mlgm)
  fmlgm  <- fixef(mlgm)
  return(list(dmlg, dmlgm, fmlg, fmlgm))
}
```

A forma de criar e passar as condições experimentais é também análoga, também em negrito as alterações em relação à binomial:

```
baseTrat <- c(-2:2)
r <- 5 # número de repetições do DIC
t <- sample(rep(baseTrat,r)) # aleatoriza os tratamentos
n <- length(t) # tamanho do experimento
m <- 3 # observações Poisson
U <- c(c(-4:0),c(-3:1),c(-2:2),c(-1:3),c(0:4)) #efeito de UE
eta <- t+U # preditor linear
lambda <- exp(eta) # media da Poisson na UE
```